Diagnostic Accuracy of Differential-Diagnosis Lists Generated by Generative Pretrained Transformer 3 Chatbot for Clinical Vignettes with Common Chief Complaints: A Pilot Study

Year: 2023 | Citations: 260 | Authors: Takanobu Hirosawa, Y. Harada, Masashi Yokose, Tetsu Sakamoto, Ren Kawamura

---

Abstract

The diagnostic accuracy of differential diagnoses generated by artificial intelligence (AI) chatbots, including the generative pretrained transformer 3 (GPT-3) chatbot (ChatGPT-3) is unknown. This study evaluated the accuracy of differential-diagnosis lists generated by ChatGPT-3 for clinical vignettes with common chief complaints. General internal medicine physicians created clinical cases, correct diagnoses, and five differential diagnoses for ten common chief complaints. The rate of correct diagnosis by ChatGPT-3 within the ten differential-diagnosis lists was 28/30 (93.3%). The rate of correct diagnosis by physicians was still superior to that by ChatGPT-3 within the five differential-diagnosis lists (98.3% vs. 83.3%, $p = 0.03$). The rate of correct diagnosis by physicians was also superior to that by ChatGPT-3 in the top diagnosis (53.3% vs. 93.3%, $p < 0.001$). The rate of consistent differential diagnoses among physicians within the ten differential-diagnosis lists generated by ChatGPT-3 was 62/88 (70.5%). In summary, this study demonstrates the high diagnostic accuracy of differential-diagnosis lists generated by ChatGPT-3 for clinical cases with common chief complaints. This suggests that AI chatbots such as ChatGPT-3 can generate a well-differentiated diagnosis list for common chief complaints. However, the order of these lists can be improved in the future.